AdaMixCon: Difficulty-Aware Feature Enhancement for Few-Shot Hyperspectral Image Classification

Chenchen Wang, Jiamu Sheng, Jingyi Zhou, *Student Member, IEEE*, Zhende Song, Jiayuan Fan, *Senior Member, IEEE*, and Zhouchen Lin, *Fellow, IEEE*

***Abstract*—In recent years, few-shot learning (FSL) has been widely researched in hyperspectral image (HSI) classification, aiming to mitigate the challenge of insufficient labeled data due to the costly annotation process. However, existing methods primarily focus on intrasample feature enhancement, which treats each sample as an independent individual, overlooking the valuable intersample dependencies inherent in HSI data, resulting in suboptimal performance. Moreover, due to the lack of awareness and modeling of sample difficulty, current frameworks often face a tradeoff between the optimal classification of hard samples and the risk of overfitting on easy ones, ultimately limiting the performance. To address the aforementioned challenges, we propose AdaMixCon, an adaptive difficulty-aware two-stage framework for few-shot hyperspectral image classification. First, we introduce HSI-MixCon, a spectral–spatial feature enhancement module that jointly exploits intersample dependencies and extracts intrasample spectral–spatial information. In addition, a novel difficulty-aware sample routing (DASR) module is designed to adaptively adjust the processing flow based on sample difficulty. The sample difficulty is estimated by performing internal assessment and mutual agreement between global and local predictions from the first stage, enabling early exit for easy samples and further refinement for hard ones. The experimental results demonstrate that our AdaMixCon outperforms state-of-the-art few-shot HSI classification methods on three public HSI datasets. The code is available at https://github.com/doctorlightt/AdaMixCon**

***Index Terms*—Adaptive, difficulty-aware, few-shot learning (FSL), hyperspectral image (HSI) classification, MixCon.**

I. Introduction

HYPERSPECTRAL imaging has become a powerful tool in Earth observation, providing rich spectral information that enables fine-grained material identification and analysis [1]. Compared to conventional RGB or multispectral images with only a limited number of broad spectral bands, hyperspectral image (HSI) acquires hundreds of contiguous narrow bands with high spectral resolution, making it possible to distinguish materials with subtle spectral differences. HSI classification is a fundamental task that assigns semantic labels to each pixel of HSI. This task is critical for accurate scene understanding and supports a wide range of downstream applications, including land use mapping [2], [3], environmental monitoring [4], [5], and urban planning [6].

Early deep learning-based HSI classification typically relies on large amounts of annotated data [7], [8], [9], while acquiring sufficient labeled samples of HSI is challenging due to the high cost and labor-intensive nature of manual annotation. With limited supervision signals, models are prone to overfitting on redundant or trivial patterns, ultimately hindering their performance. To address these issues, few-shot learning (FSL) has emerged as a promising paradigm. FSL typically follows a support-query setup, where the model learns from a small set of labeled support samples and then makes predictions on unseen query samples.

Current FSL methods for HSI classification can be broadly divided into three main categories.

1) *Data-Driven Methods:* These methods leverage additional training data to enhance model generalization, such as data augmentation [10] and semi-supervised learning [11], [12].
2) *Learning-Based Methods:* In contrast, these approaches leverage advanced learning frameworks to improve feature discriminability, such as metric learning [13], [14], and self-distillation [15], [16].
3) *Hybrid Data-Learning Methods:* These methods take advantage of both data-driven and learning-based methods, simultaneously exploiting external data resources and sophisticated learning mechanisms, such as transfer learning [17], [18], [19]. Despite recent advancements, existing approaches still face key challenges.

First, recent few-shot HSI classification methods often struggle to learn sufficiently discriminative features due to limited training samples. Thus, many works have focused on extracting and fusing spectral–spatial information within each sample [20], [21], [22]. These methods enhance feature representations using attention mechanisms [20], residual connections [21], or adaptive fusion strategies [22], which help highlight informative regions and enrich the final features. However, these approaches treat each sample as an independent entity and ignore the underlying dependencies across samples—an important source of class-relevant and comple-

Received 14 September 2025; revised 4 January 2026; accepted 30 January 2026. Date of publication 6 February 2026; date of current version 20 February 2026. This work was supported by the National Natural Science Foundation of China under Grant 62276004 and Grant 62571138. *(Corresponding author: Jiayuan Fan.)*

Chenchen Wang and Jingyi Zhou are with the College of Future Information Technology, Fudan University, Shanghai 200433, China.

Jiamu Sheng, Zhende Song, and Jiayuan Fan are with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai 200433, China (e-mail: jyfan@fudan.edu.cn).

Zhouchen Lin is with the State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Institute for Artificial Intelligence, Peking University, Beijing 100871, China.

Digital Object Identifier 10.1109/TGRS.2026.3661637

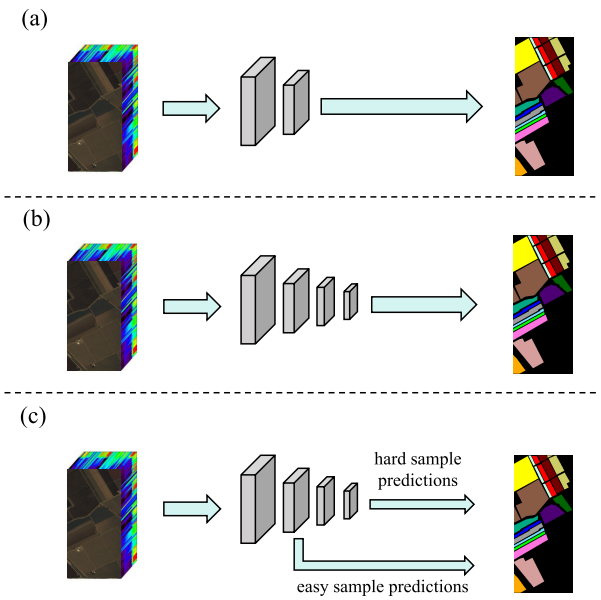

Fig. 1. Illustration of existing few-shot HSI classification frameworks. (a) Lightweight frameworks reduce overfitting but struggle with hard samples. (b) Complicated frameworks improve performance on hard samples but risk overfitting on easy ones. (c) Our proposed AdaMixCon adaptively routes samples based on difficulty, enabling early exit for easy samples and refinement for hard ones, thus achieving balanced and effective classification.

mentary information in HSI data. The lack of intersample modeling limits their capacity to fully exploit the structure of hyperspectral scenes, especially under few-shot conditions. To explore richer representations, MixCon [23] has been proposed as a sequence model combining attention and mixture of experts (MoEs), and presents a promising approach for spectral–spatial feature modeling. However, directly applying MixCon to few-shot HSI classification is limited in performance due to several limitations.

1) MixCon is designed for sequence modeling in natural language tasks, operating only on 2-D sequential inputs and lacking the capability to process visual inputs.
2) MixCon employs a unidirectional scan of language tokens during sequence modeling. However, HSI carries distinctive visual information in both spectral and spatial dimensions. MixCon's scanning method for natural language has limitations in exploiting the spectral–spatial features of HSI.
3) MixCon learns intrasample information through a unidirectional scan and attention mechanism, lacking the utilization of intersample information. Moreover, the attention mechanism of MixCon faces challenges in capturing the spectral–spatial correlations in few-shot HSI.

Therefore, a more comprehensive MixCon for few-shot HSI classification should consider the extraction of spectral–spatial features and the modeling of both intrasample and intersample dependencies.

Second, current few-shot HSI classification methods face key challenges in effectively handling samples with different difficulty levels, as depicted in Fig. 1. Lightweight models are often employed to reduce overfitting on easy samples, but they may lack the capacity to classify hard ones. Conversely, using deeper or more complicated architectures enhances performance on hard samples but increases the risk of overfitting on easy cases. Some works have explored hybrid network architectures to mitigate this dilemma. Parallel structures [24], [25] allow each branch to make independent predictions, which are subsequently fused. However, due to their relatively shallow design, each branch possesses limited representational capacity. As a result, these structures often fail to correctly classify hard samples, leading to erroneous final predictions. Other methods with cascaded architectures [26], [27] progressively refine features across layers, yet may over-process easy samples, corrupting initially correct predictions and leading to performance degradation. These limitations underscore the need for a difficulty-aware framework that can adaptively allocate processing capacity—ensuring sufficient refinement for hard samples while avoiding unnecessary complexity for easy ones.

To address the aforementioned challenges, we propose AdaMixCon, an adaptive two-stage few-shot classification framework that achieves enriched spectral–spatial modeling and enables difficulty-aware processing for HSI samples. In particular, we first employ MixCon to operate across both spectral and spatial dimensions, termed HSI-MixCon, with intersample attention and intrasample attention modules to capture rich spectral–spatial information lying under HSI data. Building upon HSI-MixCon, we design a two-stage classification pipeline. In the first stage, the model generates initial predictions, while in the second stage, an adaptive refinement is performed through a difficulty-aware sample routing (DASR) module. DASR evaluates the confidence of early-stage predictions based on internal assessment and mutual agreement and routes samples adaptively: high-confidence (easy) samples exit early, whereas low-confidence (hard) samples undergo further refinement. The two-stage pipeline enables adaptive processing for samples with different difficulty. In summary, our main contributions are as follows.

1) We propose AdaMixCon, an adaptive difficulty-aware two-stage framework for few-shot hyperspectral image classification, which enriches spectral–spatial modeling and adaptively routes each sample to two prediction streams based on their difficulty.
2) We propose HSI-MixCon, a spectral–spatial feature enhancement module, which models intersample dependencies and enhances intrasample spectral–spatial representation, enriching discriminative feature learning.
3) We propose a DASR module that evaluates the confidence of early-stage prediction based on internal assessment and mutual agreement and adaptively routes samples according to their difficulty. DASR enables effective classification for hard samples while mitigating overfitting for easy samples.

The remainder of this article is organized as follows. Section II describes related work. Section III introduces our AdaMixCon in detail. Section IV conducts extensive experiments on three HSI datasets to demonstrate the effectiveness of the proposed method. Finally, Section V draws some conclusions.

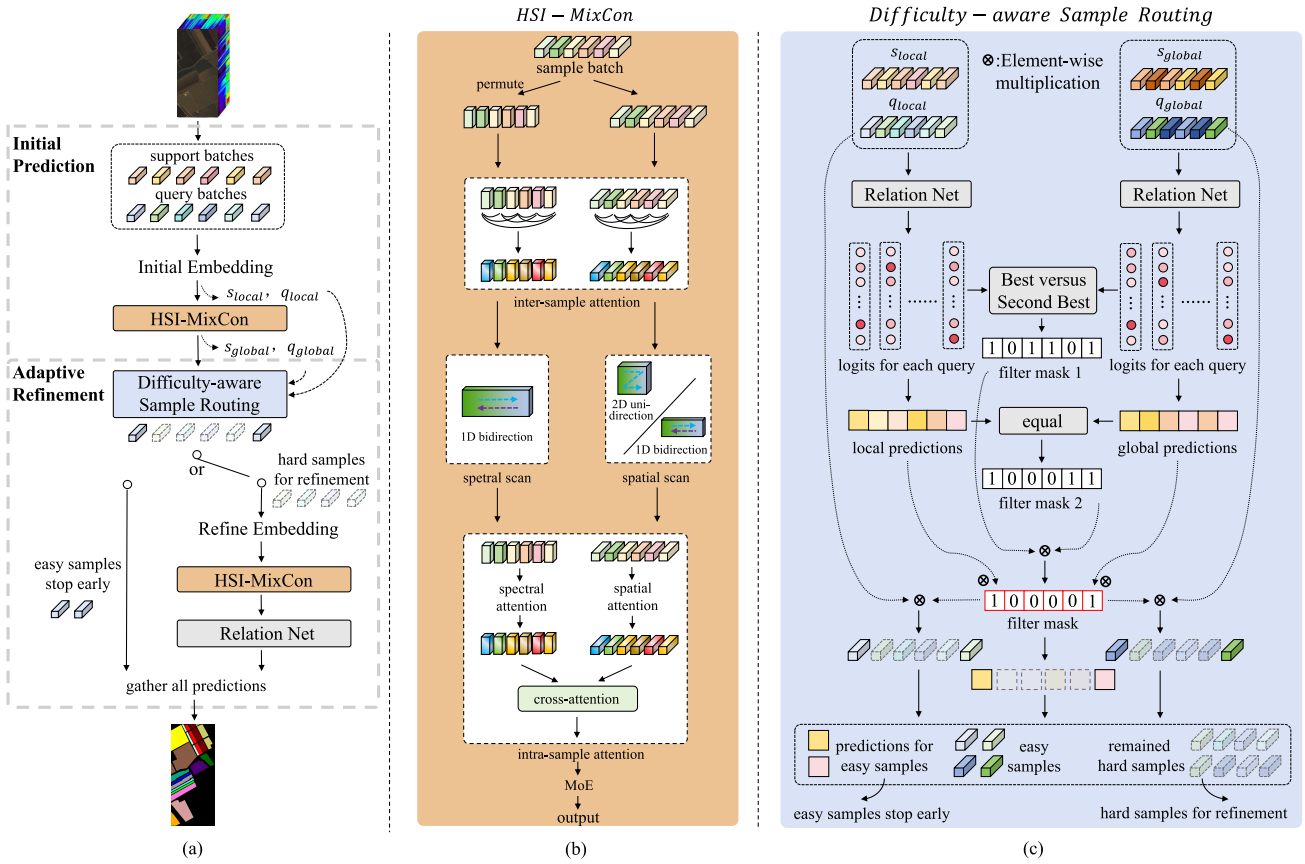

Fig. 2. Overview of our proposed AdaMixCon. The overall framework is presented in (a). AdaMixCon consists of two stages: an initial prediction stage and an adaptive refinement stage, where easy samples stop early, and only hard samples continue for refinement. The key components of AdaMixCon are HSI-MixCon and the DASR module, which are illustrated in (b) and (c), respectively. (b) HSI-MixCon combines spectral–spatial scan with intersample and intrasample attention modules to capture rich information lying under HSI samples, enabling more discriminative feature learning. (c) DASR establishes an adaptive processing flow where easy samples complete classification at the first stage, while hard samples continue for refinement, enabling effective classification for hard samples while mitigating overfitting for easy samples.

II. Related Work

A. Deep Learning Methods for HSI Classification

With the rapid development of deep learning in computer vision, deep learning-based models have become the mainstream approach for HSI classification. In particular, convolutional neural networks (CNNs) have achieved remarkable success by leveraging local receptive fields and weight sharing, effectively capturing spatial context and spectral structures. For example, Zhong et al. [28] develop a supervised spectral–spatial residual architecture, which integrates stacked 3-D convolutional operations within optimized spectral and spatial residual modules to extract highly discriminative fused features.

However, HSIs are inherently high-dimensional and contain long-range dependencies. Relying solely on local receptive fields often fails to model global relationships across distant pixels. To address this, recent works have explored global modeling techniques to enhance the overall representation and discriminative power of features. Transformers, with their self-attention mechanisms, have shown strong capability in capturing long-range dependencies and have been widely adopted in HSI classification. Qing et al. [29] introduce SAT Net, a transformer-based end-to-end model for hyperspectral classification that combines spectral and self-attention mechanisms to extract spectral–spatial features. Similarly, structured state space models (SSMs) like Mamba offer an efficient and effective alternative for sequence modeling and long-range dependency learning in HSI tasks. Li et al. [30] propose MambaHSI, which employs Mamba's superior long-range modeling to learn spatial–spectral features in HSI data through dedicated spatial and spectral Mamba blocks.

Motivated by the need to capture both fine-grained local patterns and long-range global dependencies, some recent studies seek to adopt a global–local feature extraction framework. Zhang et al. [31] introduce the convolution transformer mixer (CTMixer), which integrates CNN's local feature extraction with Transformer's global modeling through a parallel dual-branch architecture and a novel convolution-enhanced multihead attention mechanism. Sheng et al. [22] propose DualMamba, which innovatively combines lightweight CNNs' local feature extraction with Mamba's linear-complexity global modeling through a parallel dual-stream architecture, achieving efficient global–local spectral–spatial representation for HSI classification.

However, existing models mainly focus on intrasample feature enhancement, which overlooks the potential dependencies

across different samples. As a result, the learned representations may still lack precision, leading to suboptimal classification performance. Therefore, we propose intersample and intrasample attention modules to obtain enriched spectral–spatial feature representation.

B. FSL for HSI Classification

Due to the high cost of annotation and the difficulty in acquiring sufficient labeled data, FSL for HSI classification has drawn increasing attention in recent years. Existing FSL methods can be broadly categorized into three types: data-driven methods, learning-based methods, and hybrid data-learning methods.

The first category, data-driven methods, aims to enhance model generalization by introducing additional data resources. Representative strategies include data augmentation, semi-supervised learning, and so on. Wei et al. [10] propose FSHyperRGAN, a relational GAN-guided FSL framework that generates synthetic HSI samples while preserving class relationships to overcome data scarcity. Zhao et al. [11] introduce a semi-supervised framework, which progressively expands reliable training data by selectively incorporating high-confidence pseudo-labels through spatial–spectral consistency checks.

The second category, learning-based methods, focuses on improving model discriminability without relying on external data. Among these approaches, metric learning has been a classic and effective solution for few-shot HSI classification, owing to its ability to learn feature relationships from limited samples while maintaining robust discriminability. For example, Zeng et al. [25] introduce DM-MRN, a dual-metric learning framework for few-shot HSI classification, which combines image-to-class and image-to-image relation modules, enhanced by an adaptive fusion strategy to overcome single-metric limitations. In addition, drawing from model compression principles, Wu et al. [15] propose CSSD, which implements spectral-aware self-distillation by distilling central pixel spectral knowledge to guide full patch classification, resolving the granularity mismatch problem in few-shot HSI classification. Similarly, leveraging automated machine learning principles, Xiao et al. [32] introduce HCFSL-neural architecture search (NAS), which revolutionizes few-shot HSI classification by replacing manual architecture design with NAS architecture.

The third category, hybrid data-learning methods, combines external data resources and novel learning mechanisms for few-shot HSI classification. A representative approach is transfer learning. Transfer learning enables effective few-shot HSI classification by transferring learned patterns from the source domain to the target domain, requiring only minimal labeled data for adaptation. Ye et al. [17] propose GCC-FSL, a cross-domain FSL framework that leverages graph convolution contrast and prototype alignment to effectively transfer knowledge from source domains. Liu et al. [18] propose MLPA, which advances cross-domain transfer learning for few-shot HSI classification through multilevel prototype alignment and adversarial training, effectively bridging source-target domain gaps while preserving interclass discriminability in few-shot scenarios.

Although these approaches help mitigate data scarcity, current few-shot HSI classification methods still struggle to simultaneously handle samples with different difficulty levels. They either fail to classify hard samples or overfit on easy samples, ultimately limiting their performance. An ideal FSL framework for HSI classification should be capable of handling hard examples effectively while avoiding overfitting on easy ones. Therefore, we propose a difficulty-aware framework for few-shot HSI classification, which processes samples differently based on their complexity, resulting in an adaptive difficulty-aware pipeline.

III. Method

The overall architecture of our proposed AdaMixCon is depicted in Fig. 2. First, to extract enriched spectral–spatial information underlying HSI samples, we propose HSI-MixCon. Then, we build a two-stage architecture based on HSI-MixCon, where initial predictions are made in the first stage, and an adaptive refined predictions are made in the second stage. The key component of the adaptive refinement stage is a novel DASR module. We implement our framework in a support-query paradigm, which operates in a k-shot n-way FSL setup. The model is given n novel classes with k labeled examples per class, called the support set, and then classifies new query samples by comparing them with these support examples. In Sections III-A–III-C, we introduce the proposed AdaMixCon in detail.

A. Preliminary

1) State Space Model: SSMs have emerged as a powerful sequential modeling paradigm, initially revolutionizing natural language processing through efficient long-range dependency capture. The primary representation of SSMs is the continuous time representation, which transforms a time-dependent set of inputs $x(t) \in \mathbb{R}$ into a set of outputs $y(t) \in \mathbb{R}$ using a hidden state $h(t) \in \mathbb{R}^N$

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t) \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ stores all the previous history information represented by a matrix of coefficients, $\mathbf{B} \in \mathbb{R}^{N\times 1}$ is the projection parameter matrix that determines how much the input $x(t)$ affects the hidden state, $\mathbf{C} \in \mathbb{R}^{1\times N}$ denotes the state-to-output projection matrix, and $\mathbf{D} \in \mathbb{R}$ represents the direct input-to-output skip connection coefficient which is often regarded as 0. Then, the structured state space sequence model (S4) is introduced to facilitate the processing of discrete token input. After S4, selective SSM (S6) is proposed to achieve dynamic scan, which is the key component of Mamba [33].

2) MixCon for Vision Tasks: MixCon is an architecture that combines Conba layers with attention mechanisms and MoE, enabling effective global–local modeling with adaptive control. However, its potential for visual tasks remains unexplored. The key component of MixCon is Conba, which integrates a feedback mechanism with Mamba. However, Mamba cannot effectively model visual features, so we employ VMamba [34] to construct MixCon for vision tasks instead of Mamba.

In particular, given an input $\mathbf{X}_{\text{in}} \in \mathbb{R}^{B\times H\times W\times C}$, the model computes the output through the following procedure:

$$\begin{aligned}\mathbf{X}_{\text{mid}} &= \mathbf{X}_{\text{in}} + \text{SS2D}\left(\text{LayerNorm}\left(\mathbf{X}_{\text{in}}\right)\right)\\ \mathbf{X}_{\text{out}} &= \mathbf{X}_{\text{mid}} + \text{MoE}\left(\text{LayerNorm}\left(\mathbf{X}_{\text{mid}}\right)\right)\end{aligned} \tag{2}$$

where SS2D(·) denotes the 2D-selective-scan block designed for vision tasks, and MoE(·) denotes the MoE operation, which selects different multilayer perceptrons (MLPs) for feed-forward computation based on different feature distributions. As for the feedback part of Conba and the attention mechanism of MixCon, we propose new modules to make them more effective for few-shot HSI classification, and we will explain in detail in the following section. Moreover, for an input $\mathbf{X}'_{\text{in}} \in \mathbb{R}^{B\times l\times C}$ with one spatial dimension, MixCon is able to obtain $\mathbf{X}'_{\text{out}} \in \mathbb{R}^{B\times l\times C}$ through the 1-D variant of the aforementioned operations.

B. HSI-MixCon

The proposed HSI-MixCon is designed to capture rich discriminative information embedded in HSI samples from multiple perspectives, thereby enhancing few-shot HSI classification. To this end, we employ MixCon to perform spectral and spatial scans, and propose intersample and intrasample attention modules to extract rich information across and within HSI samples.

1) Intersample Attention Module: In few-shot HSI classification, limited training samples often yield incomplete or noisy features. While traditional methods focus solely on intrasample enhancement, they overlook the complementary information across samples. To this end, we propose an intersample attention module, which captures sample dependencies within each batch through similarity-based operations for feature enhancement. In this way, each sample can learn complementary information from the entire batch, strengthening the feature representation in a global manner.

We take spectral intersample attention, for example. Given a sample batch $\mathbf{X} \in \mathbb{R}^{B\times D\times H\times W}$, where $D = \text{ratio}_{\text{ssm}} * C$, and $\text{ratio}_{\text{ssm}}$ is the feature expansion factor inside SS2D, we first compute spatial average to obtain $\mathbf{X}_{\text{spec-pool}} \in \mathbb{R}^{B\times D}$, and then compute the cosine similarity between each sample pair within the batch to get pairwise similarity score $\mathbf{S}_{\cos} \in \mathbb{R}^{B\times B}$

$$\mathbf{S}_{\cos} = \left(\frac{\mathbf{X}_{\text{spec-pool}}}{\left\|\mathbf{X}_{\text{spec-pool}}\right\|_2}\right)\left(\frac{\mathbf{X}_{\text{spec-pool}}}{\left\|\mathbf{X}_{\text{spec-pool}}\right\|_2}\right)^{\top} \tag{3}$$

where $\|\cdot\|_2$ means $L2$ normalization. Therefore, $\mathbf{S}_{\cos}(i, j)$ denotes the similarity between the ith sample and the jth sample, and $\mathbf{S}_{\cos}(i, i) = 1$, located on the diagonal, $(i, j = 1, 2, \ldots, B)$. Then, for each sample, we preserve the second-highest scores, masking the highest scores on the diagonal and all the other lower redundant scores to obtain the sparse similarity attention map $\mathbf{A} \in \mathbb{R}^{B\times B}$. Finally, the attention map is applied to $\mathbf{X}_{\text{spec-pool}}$ and then added to the original input to get the final output $\mathbf{X}_{\text{spec-inter}} \in \mathbb{R}^{B\times D\times H\times W}$.

As for spatial intersample attention, the process is similar. We first compute the spectral average and reshape the feature into $\mathbf{X}_{\text{spa-pool}} \in \mathbb{R}^{B\times HW}$. Then, we utilize it to compute similarity scores over spatial dimension, and eventually obtain $\mathbf{X}_{\text{spa-inter}} \in \mathbb{R}^{B\times D\times H\times W}$ and for $\mathbf{X}' \in \mathbb{R}^{B\times D\times l}$, the process still works.

2) Spectral and Spatial Scan: After the intersample attention module, the output is then fed into the spectral and spatial scan. As for the spectral scan, the 1-D scan mechanism of MixCon naturally supports spectral processing, and we extend it to perform a bidirectional scan. In particular, for an input tensor $\mathbf{X}_{\text{spec-inter}} \in \mathbb{R}^{B\times D\times H\times W}$ with spectral channels D, we first rearrange the dimensions to $\mathbf{X}_{\text{spec-}r} \in \mathbb{R}^{B\times L\times D}$ via permute and flatten operations, where L equals $H*W$, and then perform bidirectional scan on the channel dimension. Bidirectional scan performs both forward and backward scans, then sums the outputs from both scanning directions to obtain the final result. Therefore, the preparation of a spectral scan can be formulated as follows:

$$\begin{aligned}\mathbf{X}_{\text{spec}} &= \text{Stack}\left(\left[\mathbf{X}_{\text{spec}-r}, \text{Flip}\left(\mathbf{X}_{\text{spec}-r}, -1\right)\right], 1\right)\\ \mathbf{X}_{\text{spec}-\text{proj}} &= \underset{b\,k\,d\,l,\,k\,c\,d\,\rightarrow\,b\,k\,c\,l}{\text{Einsum}}\left(\mathbf{X}_{\text{spec}}, \mathbf{W}_{\text{spec}-\text{proj}}\right)\end{aligned} \tag{4}$$

where $\mathbf{X}_{\text{spec}-\text{proj}} \in \mathbb{R}^{B\times 2\times L\times D}$, flip reverses the order of elements in a tensor along the last dimension, stack creates a new dimension and concatenates the two tensors along the first dimension to form a new tensor, and Einsum performs batch-aware matrix multiplication. Then, we flatten it into $\mathbf{X}_{\text{spec}-\text{flat}} \in \mathbb{R}^{B\times 2L\times D}$ and scan it by sequentially applying the SSM along the channel dimension $(t = 1, \ldots, D)$ and finally, we sum the outputs from both scanning directions and rearrange the dimensions to get $\mathbf{X}_{\text{spec}-\text{out}} \in \mathbb{R}^{B\times D\times L}$.

As for spatial scan, for an input tensor $\mathbf{X}_{\text{spa-inter}} \in \mathbb{R}^{B\times D\times H\times W}$, we first flatten the image into an ordered visual sequence $\mathbf{X}_{\text{spa-}r} \in \mathbb{R}^{B\times 1\times D\times L}$, where L equals $H * W$, and then apply unidirectional scan [22] for 2-D spatial dimension. Unidirectional scan performs only a forward scan to obtain the final result, which avoids redundant scanning directions that would introduce repetitive information. For input tensors with one spatial dimension, such as $\mathbf{X}'_{\text{spec-inter}} \in \mathbb{R}^{B\times D\times L}$ and $\mathbf{X}'_{\text{spa-inter}} \in \mathbb{R}^{B\times D\times L}$, the scan process still works.

3) Intrasample Attention Module: To better enhance spectral–spatial representations after spectral and spatial scans, we propose an intrasample attention module to capture spectral–spatial dependencies within individual samples, which is illustrated in Fig. 3.

First, we apply spectral and spatial attention to the scan outputs, respectively, to distinguish the importance of each scan position. Then, we achieve adaptive control of the attention map through a progressively updated error matrix, enabling continuous adjustment of attention across time steps.

Take the channel attention module, for example. For spectral scan output $\mathbf{X}_{\text{spec}-\text{out}} \in \mathbb{R}^{B\times D\times L}$ with the number of spectral channels D, we first compute the initial spectral attention $\mathbf{Att}_{\text{spec}} \in \mathbb{R}^{B\times D}$ through the original attention operating process. Then, multiply $\mathbf{Att}_{\text{spec}}$ by a learnable matrix $\mathbf{W}_{\text{adjust}}$ to obtain $\Delta\mathbf{Att}_{\text{spec}}$. Finally, the updated attention map is applied to $\mathbf{X}_{\text{spec}-\text{out}}$ to obtain $\overline{\mathbf{X}}_{\text{spec}}$ that highlights the importance of each scan position in the spectral dimension with external temporal information. The spatial attention module follows a similar rationale to the spectral attention module.

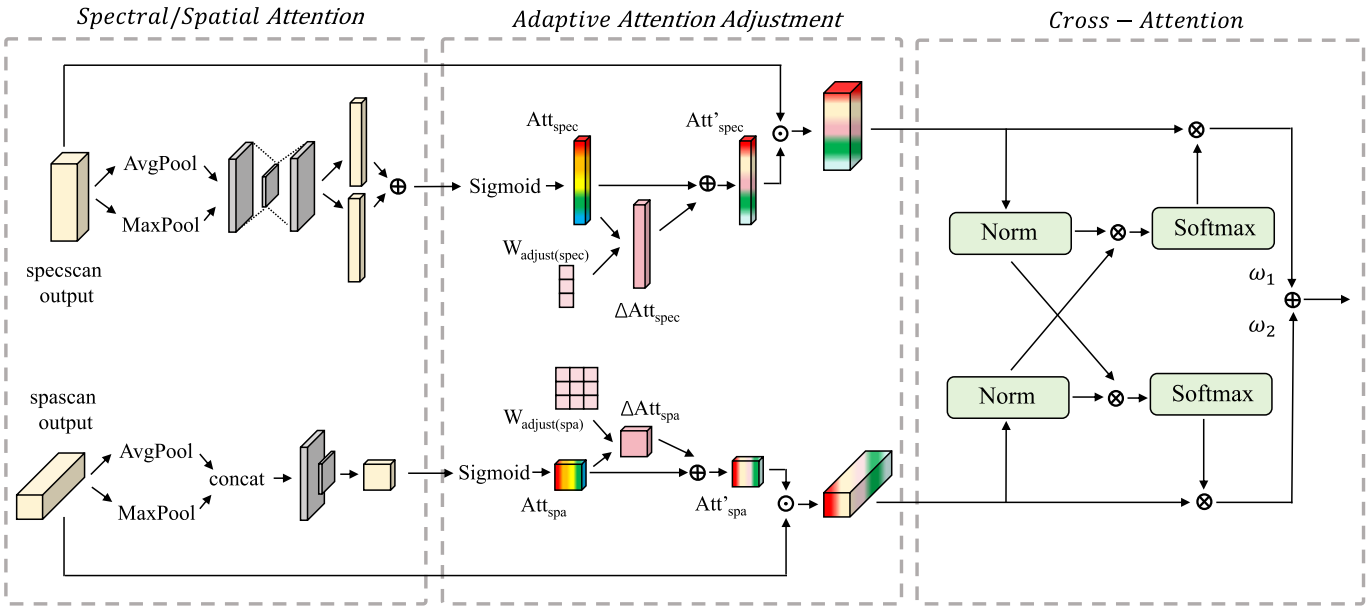

Fig. 3. Illustration of the intrasample attention module.

Second, to capture spectral–spatial dependencies, we perform spectral and spatial cross-attention operations between $\overline{\mathbf{X}}_{\text{spec}}$ and $\overline{\mathbf{X}}_{\text{spa}}$. Take spectral cross-attention for example, we first compute spectral cross-attention weights based on the similarity of the two features, and then apply it to $\overline{\mathbf{X}}_{\text{spec}}$, formulated as follows:

$$\begin{aligned}\mathbf{Sim} &= \text{Matmul}\left(\text{Norm}\left(\overline{\mathbf{X}}_{\text{spa}}\right), \text{Norm}\left(\overline{\mathbf{X}}_{\text{spec}}\right)\right)\\ \mathbf{Att}_{\text{spec-cross}} &= \text{Softmax}\left(\mathbf{Sim}\right)\\ \mathbf{X}_{\text{spec-cross}} &= \mathbf{Att}_{\text{spec-cross}}\overline{\mathbf{X}}_{\text{spec}}.\end{aligned} \tag{5}$$

Similarly, for spatial cross-attention, we first compute a spatial cross-attention map based on the similarity, and then apply it to $\overline{\mathbf{X}}_{\text{spa}}$ to obtain spatial cross-attention output $\mathbf{X}_{\text{spa-cross}}$. Finally, we compute a weighted sum of $\mathbf{X}_{\text{spec-cross}}$ and $\mathbf{X}_{\text{spa-cross}}$

$$\mathbf{X}_{\text{sum}} = \omega\mathbf{X}_{\text{spec-cross}} + (1-\omega)\mathbf{X}_{\text{spa-cross}} \tag{6}$$

where $\omega \in [0, 1]$ is the fusion weight. In this way, the model automatically learns the optimal balance between spectral and spatial representations.

C. Adaptive Difficulty-Aware Two-Stage Framework

Existing few-shot HSI classification methods face challenges in simultaneously handling samples with different difficulties. Therefore, we build an adaptive difficulty-aware two-stage framework using HSI-MixCon. The first stage produces initial predictions, while the second stage adaptively refines hard samples. The framework is detailed in three parts: the initial prediction stage, the DASR module, and the refinement process. The training and inference processes of the framework are presented in Algorithm 1.

1) Initial Prediction Stage: To fully extract spectral–spatial information from HSI data, we utilize CNN and HSI-MixCon to construct an initial prediction stage for global–local modeling. The initial prediction stage consists of data preprocessing, initial embedding, and HSI-MixCon.

As for data preprocessing, we employ principal component analysis (PCA) to reduce dimensionality and suppress redundant bands, and obtain the processed dataset $\mathcal{D}_{\text{PCA}}$. For initial embedding, we employ two convolution blocks to project the PCA-reduced data to $\mathbf{X}_{\text{local}} \in \mathbb{R}^{B\times C\times H\times W}$, where C is set to

Algorithm 1 Adaptive Difficulty-Aware Two-stage Framework

Input: Support batches $\mathbf{S}_{\text{in}}$ and query batches $\mathbf{Q}_{\text{in}}$
Output: Final predictions $\mathbf{P}_{\text{final}}$

Feed $\mathbf{S}_{\text{in}}$, $\mathbf{Q}_{\text{in}}$ into initial embedding to get $\mathbf{S}_{\text{local}}$, $\mathbf{Q}_{\text{local}}$;
Feed $\mathbf{S}_{\text{local}}$, $\mathbf{Q}_{\text{local}}$ into HSI-MixCon to get $\mathbf{S}_{\text{global}}$, $\mathbf{Q}_{\text{global}}$;
Training Stage:
for *epoch in* N_{epoch} **do**
  Feed $\mathbf{S}_{\text{local}}$, $\mathbf{Q}_{\text{local}}$, $\mathbf{S}_{\text{global}}$, $\mathbf{Q}_{\text{global}}$ into DASR to get $\mathbf{P}_{\text{easy}}$, $\mathbf{P}_{\text{local}}$, $\mathbf{P}_{\text{global}}$ and filter mask;
  Feed $\mathbf{S}_{\text{local}}$, $\mathbf{Q}_{\text{local}}$, $\mathbf{S}_{\text{global}}$, $\mathbf{Q}_{\text{global}}$ into the second stage to get $\mathbf{P}_{\text{stage2}}$, $\mathcal{L}_{MoE}$;
  Get $\mathbf{P}_{\text{hard}}$ using $\mathbf{P}_{\text{stage2}}$ and filter mask;
  Get $\mathbf{P}_{\text{final}}$ by gathering $\mathbf{P}_{\text{easy}}$ and $\mathbf{P}_{\text{hard}}$;
  Calculate $\mathcal{L}_{CE}$ using $\mathbf{P}_{\text{final}}$, $\mathbf{P}_{\text{local}}$, $\mathbf{P}_{\text{global}}$, $\mathbf{P}_{\text{stage2}}$;
  Train AdaMixCon using $\mathcal{L}_{CE}$ and $\mathcal{L}_{MoE}$ with Adam.
end
Testing Stage:
1) Feed $\mathbf{S}_{\text{local}}$, $\mathbf{Q}_{\text{local}}$, $\mathbf{S}_{\text{global}}$, $\mathbf{Q}_{\text{global}}$ into DASR to get $\mathbf{P}_{\text{easy}}$, $\mathbf{S}_{\text{hard}}$, $\mathbf{Q}_{\text{hard}}$;
2) Feed $\mathbf{S}_{\text{hard}}$, $\mathbf{Q}_{\text{hard}}$ into the second stage to get $\mathbf{P}_{\text{hard}}$;
3) Get $\mathbf{P}_{\text{final}}$ by gathering $\mathbf{P}_{\text{easy}}$ and $\mathbf{P}_{\text{hard}}$.

64. The process can be formulated as

$$\begin{aligned}\mathbf{X}_1 &= \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv2d}\left(\mathbf{X}\right)\right)\right)\\ \mathbf{X}_{\mathrm{local}} &= \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv2d}\left(\mathbf{X}_1\right)\right)\right)\end{aligned} \tag{7}$$

where BN(·) denotes BatchNorm2d operation, and Conv2d denotes 2-D convolution with kernel size 1×1.

Then, we send $\mathbf{X}_{\mathrm{local}}$ into HSI-MixCon for global modeling, and obtain $\mathbf{X}_{\mathrm{global}}$, which consists of $\mathbf{s}_{\mathrm{global}}$ and $\mathbf{q}_{\mathrm{global}}$. Subsequently, both $\mathbf{X}_{\mathrm{local}}$ and $\mathbf{X}_{\mathrm{global}}$ are passed to the second stage for adaptive refinement.

2) Difficulty-Aware Sample Routing: The output of the initial prediction stage already exhibits certain classification capability, but there are still a number of samples that fail to be correctly classified, calling for further feature refinement. However, refining all outputs from the initial prediction stage may lead to overfitting on easy samples that are already correctly classified, causing feature distortion and ultimately degrading their classification performance. To address this, we propose a DASR module.

First, we employ relation network [35] to compute the metric of $\mathbf{X}_{\mathrm{local}}$ and $\mathbf{X}_{\mathrm{global}}$. Relation network outputs a similarity score $\mathbf{S} \in \mathbb{R}^{B_q \times B_s}$ for query-support sample pairs, quantifying their pairwise similarity, where B_q and B_s denote the batch size of query and support samples. Since B_s is the product of the number of classes n and the number of samples per class k, $\mathbf{S}$ can be reshaped to $\mathbf{S} \in \mathbb{R}^{B_q \times n \times k}$. For a given query sample, we average its similarity scores with all k support samples from the ith class, and take the mean score as its probability of belonging to the ith class ($i = 1, \ldots, n$)

$$\mathbf{L}_i = \frac{1}{k}\sum_{j=1}^{k}\mathbf{S}\left(i, :, j\right), \quad i \in \left\{1, \ldots, B_q\right\} \tag{8}$$

where $\mathbf{L} \in \mathbb{R}^{B_q \times n}$ represents the relation logits for queries, and $\mathbf{L}_i \in \mathbb{R}^n$ denotes the probabilities of belonging to each class for ith query sample. Then, the class with the maximum probability in $\mathbf{L}_i$ is taken as the prediction result for the ith query sample, and finally, we obtain the initial prediction results of the query batch, termed $\mathbf{Y} \in \mathbb{R}^{B_q}$. Therefore, the prediction results of $\mathbf{X}_{\mathrm{local}}$ and $\mathbf{X}_{\mathrm{global}}$ are $\mathbf{Y}_{\mathrm{local}}$ and $\mathbf{Y}_{\mathrm{global}}$.

Second, in order to establish a reliable strategy to distinguish between easy and hard samples, we utilize two joint criteria to filter easy and hard samples.

1) Best versus second best (BvSB) for internal assessment. BvSB operates on the key insight that a sample's prediction certainty can be quantified by the margin between the highest (best) and second-highest (second best) softmax scores in the model's output probability distribution. Given a predicted probability vector $\mathbf{L}_i$ with $p_1 > \cdots > p_n$ being the probabilities sorted in descending order, BvSB is computed by

$$\mathbf{BvSB} = p_1 - p_2. \tag{9}$$

Then we compare the BvSB score with a predefined confidence threshold θ. Samples with BvSB $> \theta$ are considered to have high prediction confidence, while those with BvSB $< \theta$ are seen as low-confidence predictions.

2) Equal criterion for mutual agreement. Although BvSB can effectively identify samples with high classification confidence, it may still yield highly confident but incorrect predictions. To enhance reliability, we introduce an equal verification mechanism between $\mathbf{Y}_{\mathrm{local}}$ and $\mathbf{Y}_{\mathrm{global}}$. In particular, when the classification results derived from local features and global features are the same

$$\mathbf{Y}^i_{\mathrm{local}} = \mathbf{Y}^i_{\mathrm{global}} \quad i \in \left\{1, \ldots, B_q\right\}. \tag{10}$$

We consider such a prediction to be more trustworthy than that based solely on local or global features, and in this case, the ith sample is considered more likely to be an easy sample.

Finally, samples with BvSB $> \theta$ and mutual agreement are treated as high-confidence (easy) samples, calling for early stopping, and the remaining samples are treated as low-confidence (hard) samples, requiring further refinement.

We employ filter masks for sample routing based on the BvSB and equal criteria. In particular, samples satisfying BvSB $> \theta$ are assigned value 1, while the others are assigned 0, forming filter mask 1. Similarly, samples meeting the equal criterion are assigned 1, with the rest 0, forming filter mask 2. The final filter mask is obtained by elementwise multiplication of filter masks 1 and 2. This resulting mask is then applied to the features obtained from the first stage: features corresponding to mask value 1 are identified as easy samples and undergo early stopping, while those corresponding to 0 are treated as hard samples and proceed to the second stage for refinement.

3) Refinement for Hard Samples: The low-confidence predictions of hard samples reveal the limitations of the initial stage in extracting discriminative information from challenging cases. Therefore, we propose the refinement for hard sample features.

To be specific, first, we extend the previously mentioned initial embedding by adding a maxpool layer after each convolution block. Therefore, the hard sample features coming from DASR will undergo the following feature extraction process:

$$\begin{aligned}\mathbf{X}_1 &= \mathrm{MP}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv2d}\left(\mathbf{X}_{\mathrm{hard}}\right)\right)\right)\right)\\ \mathbf{X}_{\mathrm{conv}} &= \mathrm{MP}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv2d}\left(\mathbf{X}_1\right)\right)\right)\right)\end{aligned} \tag{11}$$

where BN(·) denotes BatchNorm2d operation, MP(·) denotes Maxpool2d operation with kernel size 2×2 and stride 2, and Conv2d denotes 2-D convolution with kernel size 1×1.

Then, we employ the image-to-class block [42] on the support and query sets to obtain the similarity features, which are fed into HSI-MixCon for sophisticated modeling to derive more reliable similarity measurements, termed the metric computing process. However, the output will be fed into the relation network to obtain the final refined predictions. Finally, we gather the predictions from the initial prediction stage and the refinement stage to generate the final predictions of all samples.

IV. Experiments and Results

In this section, we first describe three well-known public HSI datasets: the Pavia University dataset, the Salinas

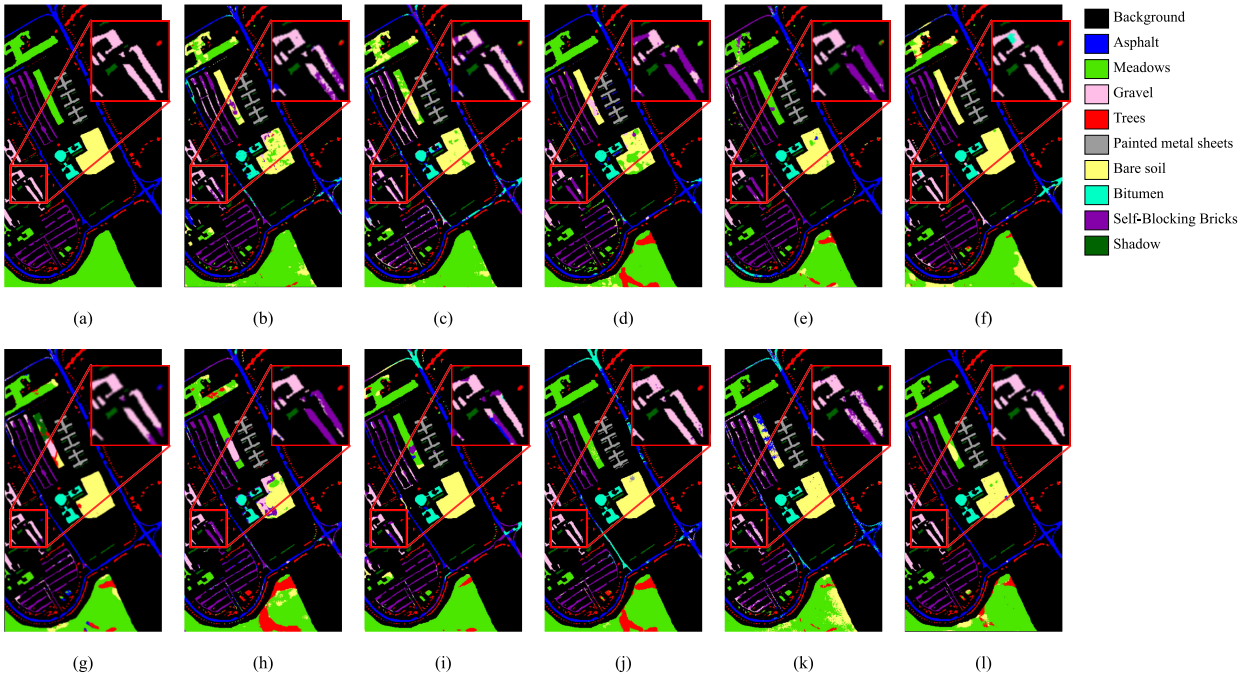

Fig. 4. Classification maps obtained by different methods on the Pavia University dataset. (a) Ground Truth. (b) S-DMM (OA = 76.67%). (c) DCFSL (OA = 79.67%). (d) S3Net (OA = 76.62%). (e) DM-MRN (OA = 87.06%). (f) FSCF-SSL (OA = 89.15%). (g) SPFormer (OA = 83.93%). (h) CTF-SSCL (OA = 82.29%). (i) MambaHSI+ (OA = 86.11%). (j) CGL-PNM (OA = 88.43%). (k) SSEFN (OA = 85.55%). (l) AdaMixCon (OA = 90.74%).

TABLE I

LAND-COVER TYPES, THE NUMBER OF LABELED TRAINING SAMPLES AND TESTING SAMPLES OF THE PAVIA UNIVERSITY DATASET

Class	Land Cover Type	Training	Testing
1	Asphalt	5	6626
2	Meadows	5	18644
3	Gravel	5	2094
4	Trees	5	3059
5	Painted metal sheets	5	1340
6	Bare soil	5	5024
7	Bitumen	5	1325
8	Self-Blocking Bricks	5	3677
9	Shadow	5	942
	Total	45	42731

TABLE II

LAND-COVER TYPES, THE NUMBER OF LABELED TRAINING SAMPLES AND TESTING SAMPLES OF THE SALINAS DATASET

Class	Land Cover Type	Training	Testing
1	Brocoli green weeds 1	5	2004
2	Brocoli green weeds 2	5	3721
3	Fallow	5	1971
4	Fallow rough plow	5	1389
5	Fallow smooth	5	2673
6	Stubble	5	3954
7	Celery	5	3574
8	Grapes untrained	5	11266
9	Soil vinyard develop	5	6198
10	Corn senesced green weeds	5	3273
11	Lettuce romaine 4wk	5	1063
12	Lettuce romaine 5wk	5	1922
13	Lettuce romaine 6wk	5	911
14	Lettuce romaine 7wk	5	1065
15	Vinyard untrained	5	7263
16	Vinyard vertical trellis	5	1802
	Total	80	54049

dataset, and the Houston 2013 dataset. We then introduce the experimental setting, including evaluation metrics and implementation details. Following this, we undertake quantitative experiments and perform an ablation analysis to assess the efficacy of our proposed method.

A. Datasets Description

1) Pavia University: The Pavia University dataset, acquired by the Reflective Optics Spectrographic Imaging System over Pavia, Italy, consists of 610 × 340 pixels with a spatial resolution of 1.3 m per pixel. Each pixel contains 103 spectral bands covering the wavelength range of 430–860 nm. The dataset contains 42 776 labeled pixels categorized into nine distinct land-cover classes. Following common practice, we utilize five labeled samples for training and reserve the others for testing. The specific class names along with their corresponding training and testing sample counts are detailed in Table I.

2) Salinas: The Salinas dataset, captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over Salinas Valley, comprises 512 × 217 pixels with an exceptional spatial resolution of 3.7 m per pixel. Each pixel contains 224 spectral bands covering the 360–2500-nm wavelength range. The dataset contains 54 129 labeled pixels distributed across 16 categories, including various crop types and soils. We adopt five labeled samples for training and the rest for testing, with detailed class distributions provided in Table II.

TABLE III
LAND-COVER TYPES, THE NUMBER OF LABELED TRAINING SAMPLES AND TESTING SAMPLES OF THE HOUSTON 2013 DATASET

Class	Land Cover Type	Training	Testing
1	Healthy grass	5	1246
2	Stressed grass	5	1249
3	Synthetic grass	5	692
4	Trees	5	1239
5	Soil	5	1237
6	Water	5	320
7	Residential	5	1263
8	Commercial	5	1239
9	Road	5	1247
10	Highway	5	1222
11	Railway	5	1230
12	Parking Lot 1	5	1228
13	Parking Lot 2	5	464
14	Tennis Court	5	423
15	Running Track	5	655
	Total	75	14954

3) Houston 2013: The Houston 2013 dataset, acquired by the ITRES CASI-1500 hyperspectral sensor over the University of Houston campus and its surrounding urban area, consists of 349 × 1905 pixels with a spatial resolution of 2.5 m per pixel. It covers 144 spectral bands in the 350–1050-nm wavelength range. The dataset contains 15 029 labeled pixels distributed across 15 urban land-cover classes. We use five labeled samples for training and the others for testing, with detailed class statistics provided in Table III.

B. Experimental Setting

1) Evaluation Metrics: We evaluate the performance of all methods by three widely used indexes: overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ).

2) Comparison With State-of-the-Art Methods: To demonstrate the effectiveness of our proposed method, we compare our classification performance with several SOTA approaches using the most effective setting for these methods. In particular, we selected two methods that utilize additional training data: SPFormer [27] and CGL-PNM [41], which employ self-supervised learners; five methods with novel learning frameworks: DM-MRN [25], S3Net [38], and SSEFN [24], CTF-SSCL [26], MambaHSI+ [40]; and three transfer learning methods: S-DMM [36], DCFSL [37], and FSCF-SSL [39].

3) Implementation Details: The proposed AdaMixCon is implemented with the PyTorch framework. The patch size is set to 25 × 25 for the Pavia University dataset, 29 × 29 for the Salinas dataset, and 15 × 15 for the Houston 2013 dataset. The number of PCA dimensions is 10, 30, and 15, respectively. The Adam optimizer is adopted with a learning rate of 1e-3, training for 1000 epochs, and a weight decay of 5e-4. We adopt the StepLR as our training scheduler, in which the learning rate is multiplied by a gamma factor of 0.1 every 500 epochs. The cross-entropy loss is used for classification. We calculate the results fairly by averaging the results of ten repeated experiments with five labeled samples per class for training. All experiments are implemented on a single NVIDIA A100 GPU, which has 80-GB memory.

C. Quantitative Results and Analysis

1) Classification Results Compared With SOTA Methods: Quantitative results of different methods in terms of class-specific accuracy, OA, AA, and Kappa coefficient on the Pavia University, Salinas, and Houston 2013 datasets are presented in Tables IV–VI, respectively, while their corresponding classification maps are shown in Figs. 4–6.

Table IV presents the experimental results on the Pavia University dataset. Our proposed AdaMixCon improves OA, AA, and Kappa by 1.59%, 1.34%, and 2.06%, respectively, compared to the second-best FSCF-SSL method. Moreover, it achieves the best classification performance in the 3rd and 6th categories (i.e., gravel and bare soil). Notably, other methods performed poorly in the gravel category, with the highest accuracy reaching only 81.05%. In contrast, our approach achieves 88.83% accuracy. Moreover, for categories where other methods already achieve high accuracy, our approach still attains exceptional accuracy levels. This demonstrates our method's successful classification of hard samples while maintaining its effectiveness on easy cases, proving our framework's unified capability in handling both simple and difficult classification tasks.

Table V presents the experimental results on the Salinas dataset. Our proposed AdaMixCon improves OA, AA, and Kappa by 2.47%, 0.62%, and 2.73%, respectively, compared to the second-best SPFormer method. Moreover, it achieves the best classification performance in the 3rd, 8th, and 10th categories. Notably, compared to other categories, other methods achieve relatively lower accuracies on the 8th category, which means samples from this category are more difficult to classify. Our method attains the highest accuracy on this category, and maintains competitive accuracy across other categories, demonstrating its capability in both enhancing hard sample classification and preserving performance on easy samples.

Table VI presents the experimental results on the Houston 2013 dataset. Our proposed AdaMixCon improves OA, AA, and Kappa by 1.73%, 1.35%, and 1.86%, respectively, compared to the second-best DM-MRN method. It achieves the best classification performance in the 9th and 14th categories. Moreover, for categories where other methods already achieved high accuracy, our approach still attains exceptional precision levels. This demonstrates our method's successful classification of hard samples while maintaining its effectiveness on easy cases.

2) Model Complexity: Table VII displays training time, testing time, and parameters of the compared methods and our AdaMixCon. Nearly half of the parameters in our model come from the MoE architecture, which enables the model to maintain a massive parameter scale for knowledge storage while activating only a few experts per input during inference. Therefore, it maintains exceptional performance without increasing inference latency. Overall, our method is efficient in terms of both running time and parameters compared to other approaches, while achieving the best performance. For example, our approach demonstrates significantly faster training time than FSCF-SSL across all three datasets, while achieving shorter testing time than DM-MRN and CGL-PNM.

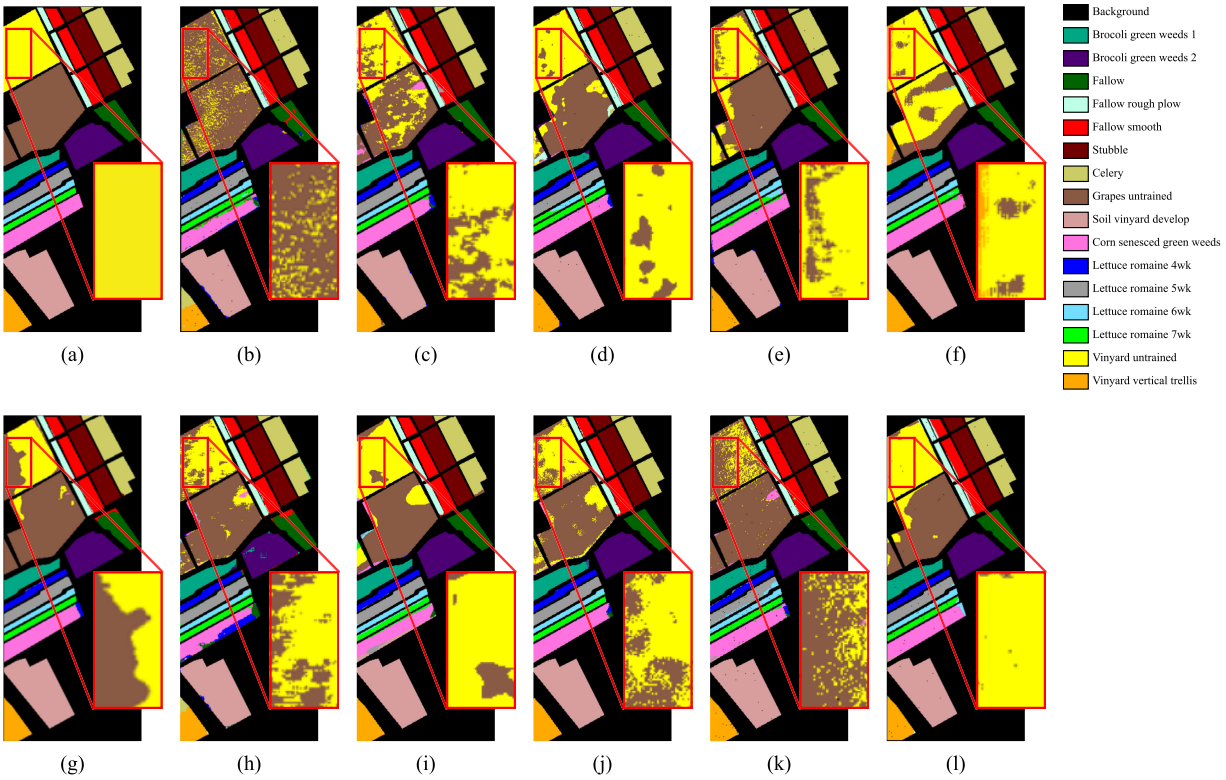

Fig. 5. Classification maps obtained by different methods on the Salinas dataset. (a) Ground Truth. (b) S-DMM (OA = 86.42%). (c) DCFSL (OA = 88.89%). (d) S3Net (OA = 92.02%). (e) DM-MRN (OA = 93.20%). (f) FSCF-SSL (OA = 91.18%). (g) SPFormer (OA = 93.22%). (h) CTF-SSCL (OA = 90.77%). (i) MambaHSI+ (OA = 91.55%). (j) CGL-PNM (OA = 90.59%). (k) SSEFN (OA = 91.20%). (l) AdaMixCon (OA = 95.69%).

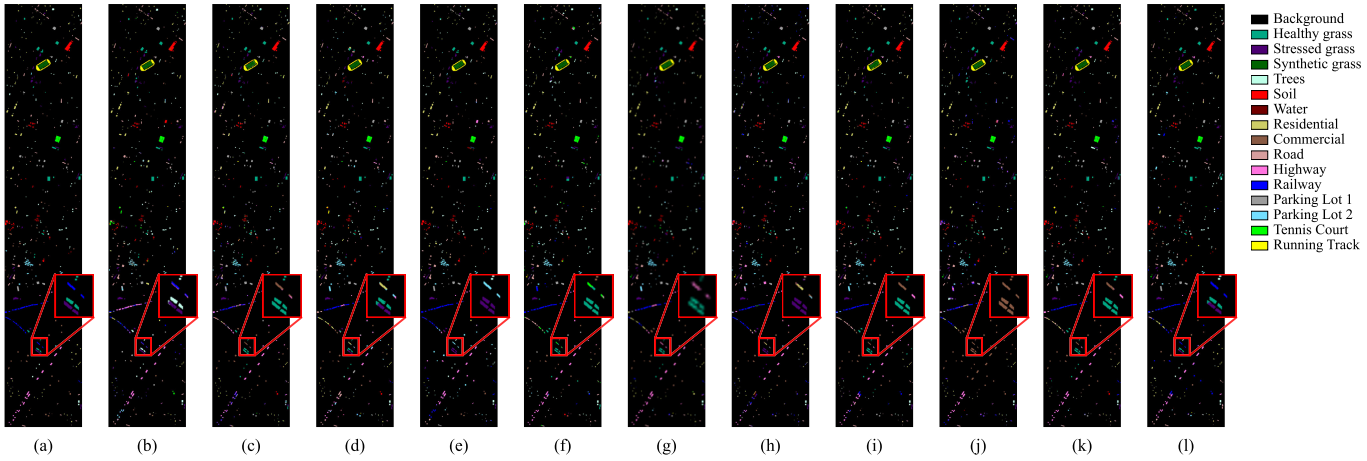

Fig. 6. Classification maps obtained by different methods on the Houston 2013 dataset. (a) Ground Truth. (b) S-DMM (OA = 73.94%). (c) DCFSL (OA = 74.68%). (d) S3Net (OA = 73.44%). (e) DM-MRN (OA = 81.42%). (f) FSCF-SSL (OA = 78.66%). (g) SPFormer (OA = 79.03%). (h) CTF-SSCL (OA = 79.85%). (i) MambaHSI+ (OA = 77.86%). (j) CGL-PNM (OA = 75.52%). (k) SSEFN (OA = 77.25%). (l) AdaMixCon (OA = 83.15%).

While our method contains more parameters than SPFormer, it substantially outperforms SPFormer in terms of accuracy, achieving 6.81%, 2.47%, and 4.12% higher OA on Pavia University, Salinas, and Houston 2013, respectively. Furthermore, compared with SSEFN, our method requires considerably fewer parameters across all three datasets.

D. Ablation Study and Analysis

1) Ablation for AdaMixCon Architecture: Our proposed AdaMixCon has two stages. In the first stage, the model generates initial predictions, while in the second stage, an adaptive refinement is performed through the DASR module.

For the first two setups in Table VIII, we use only the initial stage or refinement stage for classification, and the classification performance drops significantly. This demonstrates that using either stage alone for classification still lacks certain discriminative capabilities of hard samples. The third setup employs a cascaded two-stage framework without DASR, showing improvement over the first two setups. This

TABLE IV

QUANTITATIVE PERFORMANCE OF DIFFERENT CLASSIFICATION METHODS IN TERMS OF OA, AA, AND κ AS WELL AS THE ACCURACIES FOR EACH CLASS ON THE PAVIA UNIVERSITY DATASET. THE BEST RESULTS ARE SHOWN IN BOLD. (FIVE LABELED SAMPLES PER CLASS)

Class	S-DMM [36]	DCFSL [37]	S3Net [38]	DM-MRN [25]	FSCF-SSL [39]	SPFormer [27]	CTF-SSCL [26]	MambaHSI+ [40]	CGL-PNM [41]	SSEFN [24]	AdaMixCon
1	86.87±5.71	78.65±13.44	76.44±22.30	86.85±7.38	82.26±14.76	87.69±3.85	88.73±6.07	83.73±8.65	76.53±7.52	**97.26±2.29**	87.27±5.93
2	66.79±10.13	83.32±9.77	72.77±15.78	84.42±7.70	88.30±5.84	74.59±9.04	80.07±8.61	85.36±10.21	90.62±3.14	**97.55±1.66**	90.06±6.00
3	76.62±9.68	64.44±10.55	71.56±19.04	80.46±11.65	79.47±12.61	81.03±11.86	66.33±12.57	69.55±10.12	81.05±10.72	69.83±11.64	**88.83±8.96**
4	86.01±10.40	93.55±3.52	89.88±5.45	90.38±6.12	**97.15±1.69**	91.55±5.03	87.32±5.02	89.99±3.68	94.17±2.07	74.95±17.47	86.68±6.14
5	99.81±0.51	98.75±1.41	99.84±0.33	99.72±0.73	99.95±0.13	**100.00±0.00**	99.78±0.36	96.50±4.44	99.79±0.20	99.63±0.92	99.94±0.12
6	74.36±11.51	70.71±10.55	69.11±14.84	88.52±14.12	90.50±11.63	93.95±9.19	76.52±12.01	90.11±15.85	94.10±9.02	78.69±17.43	**94.92±5.98**
7	96.31±2.65	80.14±6.03	95.40±7.09	98.84±1.67	89.34±10.91	**99.56±0.35**	90.32±8.27	85.89±15.03	95.64±3.56	68.66±18.81	99.01±1.23
8	82.31±13.93	60.27±14.84	79.39±19.09	87.90±9.48	**96.06±5.87**	91.56±3.88	82.04±9.52	89.63±4.38	81.92±12.48	77.14±3.39	91.32±6.10
9	**99.98±0.04**	99.12±0.86	92.14±5.94	98.85±1.88	99.94±0.10	96.25±4.67	95.65±5.74	92.56±8.55	95.27±7.58	90.65±4.27	97.02±4.16
OA (%)	76.67±4.42	79.67±4.21	76.62±7.43	87.06±2.07	89.15±3.58	83.93±3.13	82.29±2.45	86.11±3.81	88.43±1.84	85.55±4.80	**90.74±2.40**
AA (%)	85.45±3.00	80.99±1.54	82.95±3.25	90.66±1.67	91.44±2.78	90.69±1.82	85.20±1.59	87.04±2.50	89.90±1.85	83.82±2.82	**92.78±1.87**
$\kappa \times 100$	70.71±5.12	73.79±4.77	70.61±8.47	83.33±2.43	85.93±4.49	79.82±3.63	77.24±2.73	85.50±4.68	84.98±2.41	81.55±5.84	**87.99±2.98**

TABLE V

QUANTITATIVE PERFORMANCE OF DIFFERENT CLASSIFICATION METHODS IN TERMS OF OA, AA, AND κ AS WELL AS THE ACCURACIES FOR EACH CLASS ON THE SALINAS DATASET. THE BEST RESULTS ARE SHOWN IN BOLD. (FIVE LABELED SAMPLES PER CLASS)

Class	S-DMM [36]	DCFSL [37]	S3Net [38]	DM-MRN [25]	FSCF-SSL [39]	SPFormer [27]	CTF-SSCL [26]	MambaHSI+ [40]	CGL-PNM [41]	SSEFN [24]	AdaMixCon
1	99.44±0.54	99.39±0.92	98.25±2.81	99.64±0.47	99.78±0.66	99.87±0.33	99.36±1.10	99.25±1.11	98.15±1.77	**100.00±0.00**	99.18±1.51
2	98.94±1.29	99.48±0.99	97.99±3.03	99.85±0.40	98.83±1.88	**100.00±0.00**	99.34±1.14	99.17±1.95	99.58±0.62	99.53±0.68	99.68±0.79
3	96.94±3.91	91.43±10.57	99.29±2.13	99.99±0.03	95.07±7.26	94.93±5.86	92.37±9.68	99.19±1.63	93.94±10.08	97.28±2.42	**99.99±0.01**
4	99.23±1.05	99.50±0.53	96.05±10.12	98.48±2.16	**99.89±0.14**	99.86±0.30	99.40±0.61	97.81±2.74	99.11±1.09	96.83±1.33	98.55±3.12
5	96.74±4.43	90.83±4.48	97.36±2.81	93.89±5.21	96.67±1.74	95.98±3.62	93.86±3.40	95.97±3.86	89.45±11.19	**97.61±1.48**	94.31±4.75
6	99.18±0.83	98.28±1.89	98.27±2.79	98.98±1.88	99.69±0.41	99.33±0.84	99.27±0.97	98.11±2.52	97.72±2.50	**99.99±0.01**	99.51±1.26
7	99.70±0.43	99.12±1.03	99.82±0.24	99.00±1.27	97.25±6.97	99.90±0.17	99.66±0.38	92.18±19.72	99.73±0.41	**99.91±0.15**	99.70±0.44
8	61.53±15.19	72.93±13.78	77.51±14.01	82.38±10.85	80.60±8.56	81.95±8.02	82.80±4.14	71.49±16.27	76.40±10.45	83.46±5.38	**89.52±5.44**
9	98.78±1.02	99.36±0.79	98.47±3.10	98.88±1.88	99.22±0.78	**100.00±0.00**	98.98±0.96	99.91±0.15	99.75±0.25	99.19±0.35	99.99±0.02
10	84.10±8.35	85.61±7.38	89.97±4.93	92.52±5.54	93.64±5.99	92.77±4.67	84.39±9.41	91.66±10.59	92.82±4.72	95.37±1.38	**95.43±4.51**
11	96.03±3.49	98.13±2.26	99.21±0.84	98.97±2.02	99.53±0.46	**99.57±1.12**	94.45±5.28	99.41±0.71	99.45±0.65	94.69±2.55	98.72±2.85
12	**99.85±0.27**	99.28±0.96	94.51±4.13	96.70±3.98	99.19±1.30	97.73±2.82	99.13±0.74	99.26±1.06	97.83±2.09	99.41±0.39	98.26±2.09
13	**99.79±0.28**	99.34±0.49	95.85±3.92	94.92±6.72	99.51±0.51	98.28±1.16	98.93±1.20	99.16±0.80	98.08±1.32	97.24±3.85	96.38±4.55
14	93.80±4.11	98.25±2.18	94.24±5.86	90.36±10.04	97.08±3.49	98.32±3.76	98.51±1.39	97.40±4.31	**99.17±1.14**	92.50±5.61	90.07±10.48
15	74.52±21.69	77.10±7.64	89.87±7.01	88.10±8.33	73.90±14.49	86.38±8.11	74.62±7.79	**94.80±3.79**	79.13±8.47	70.89±6.58	92.25±8.62
16	89.28±7.48	90.57±6.21	95.29±5.47	99.13±1.17	98.51±1.18	95.61±4.31	92.85±5.27	95.25±4.67	98.87±1.46	**99.97±0.05**	98.81±2.83
OA (%)	86.42±3.02	88.89±2.52	92.02±2.68	93.20±1.81	91.18±1.41	93.22±1.55	90.77±1.39	91.55±3.29	90.59±2.13	91.20±1.03	**95.69±1.00**
AA (%)	92.99±1.59	93.66±1.22	95.12±1.40	95.74±0.74	95.52±1.02	96.28±0.75	94.24±1.36	95.63±1.57	94.96±1.54	95.24±0.43	**96.90±0.72**
$\kappa \times 100$	84.94±3.35	87.67±2.76	91.15±2.94	92.45±1.99	90.19±1.57	92.47±1.72	89.74±1.55	94.93±1.37	89.55±2.34	90.21±1.15	**95.20±1.11**

TABLE VI

QUANTITATIVE PERFORMANCE OF DIFFERENT CLASSIFICATION METHODS IN TERMS OF OA, AA, AND κ AS WELL AS THE ACCURACIES FOR EACH CLASS ON THE HOUSTON 2013 DATASET. THE BEST RESULTS ARE SHOWN IN BOLD. (FIVE LABELED SAMPLES PER CLASS)

Class	S-DMM [36]	DCFSL [37]	S3Net [38]	DM-MRN [25]	FSCF-SSL [39]	SPFormer [27]	CTF-SSCL [26]	MambaHSI+ [40]	CGL-PNM [41]	SSEFN [24]	AdaMixCon
1	89.80±7.07	89.69±7.28	76.30±12.36	77.32±11.41	72.66±13.52	84.45±3.78	**89.95±6.74**	85.84±6.19	79.87±4.03	87.83±6.89	85.12±4.41
2	85.20±11.14	81.94±4.91	85.73±9.25	**89.14±10.87**	62.76±12.73	73.63±14.51	86.43±8.12	77.11±6.00	69.28±14.41	88.78±9.95	86.83±10.22
3	99.49±0.55	95.34±4.18	98.84±1.08	98.67±1.27	99.29±0.74	99.32±0.99	96.15±5.10	93.99±8.50	96.86±1.99	**99.54±0.31**	98.46±1.55
4	**96.10±2.70**	91.86±3.59	88.90±6.28	92.23±5.68	88.09±8.92	93.15±4.11	93.89±4.30	87.08±5.19	75.58±11.71	85.12±14.89	92.69±5.12
5	96.33±6.27	96.41±3.36	90.26±8.75	98.45±4.04	91.21±8.11	99.35±1.00	95.89±6.39	**99.49±1.26**	94.79±5.84	88.06±5.05	97.83±5.33
6	89.19±4.36	78.32±6.64	94.10±5.07	90.18±7.39	91.50±4.82	91.08±5.78	85.84±7.65	77.56±12.49	88.86±3.74	**96.51±8.06**	89.55±7.66
7	73.36±12.26	58.25±10.07	62.03±9.16	81.29±12.84	79.48±8.80	**84.14±9.07**	70.40±12.73	72.81±13.90	82.68±6.15	77.24±8.44	82.81±9.90
8	40.35±6.97	48.30±9.30	42.47±8.10	41.93±8.82	44.97±14.45	42.10±8.62	50.20±11.74	41.90±9.23	57.84±9.01	**72.16±17.48**	51.91±12.58
9	67.84±15.94	51.58±13.39	58.99±9.10	76.37±3.62	76.04±6.25	75.66±10.58	68.85±8.28	72.38±11.07	64.92±12.02	68.45±10.38	**77.17±5.16**
10	48.82±14.27	63.74±12.19	66.76±13.42	75.25±11.78	**78.87±15.82**	71.74±10.69	71.14±11.96	73.85±8.66	63.03±11.37	68.85±11.48	75.43±10.42
11	63.63±12.82	61.51±12.02	66.13±13.48	84.23±9.30	**87.44±9.65**	78.53±7.55	68.51±11.78	73.78±12.52	70.47±10.79	71.56±13.04	82.43±8.01
12	51.80±14.21	**77.24±14.95**	52.72±11.92	64.76±8.65	67.61±11.13	52.10±9.72	75.74±13.23	69.45±14.72	58.19±11.82	65.56±12.51	71.82±7.84
13	50.40±21.47	82.65±8.30	84.91±11.75	**93.92±2.38**	91.15±4.10	89.11±4.68	91.03±4.75	90.41±2.35	90.46±8.44	63.07±14.26	92.63±5.56
14	99.74±0.39	90.14±5.79	99.88±0.18	99.96±0.08	98.44±4.67	99.90±0.13	98.05±1.88	98.29±3.45	99.45±0.46	96.82±2.93	**99.96±0.12**
15	99.74±0.30	95.89±3.43	99.65±0.86	99.45±1.50	97.16±2.88	**99.74±0.21**	97.44±2.64	88.67±12.32	94.85±3.54	92.78±3.29	98.76±2.83
OA (%)	73.94±2.19	74.68±1.37	73.44±1.93	81.42±2.35	78.66±2.72	79.03±1.70	79.85±1.63	77.86±2.59	75.52±1.76	77.25±2.19	**83.15±1.86**
AA (%)	76.79±2.01	77.52±1.36	77.84±1.68	84.21±2.06	81.78±2.32	82.27±1.50	82.63±1.00	80.17±2.43	79.14±1.61	81.49±2.42	**85.56±1.79**
$\kappa \times 100$	71.88±2.35	72.66±1.47	71.34±2.08	79.95±2.53	76.95±2.94	77.38±1.84	78.24±1.75	80.00±4.40	73.58±1.91	75.42±2.35	**81.81±2.01**

TABLE VII

RUNNING TIME (S) AND PARAMETER SIZE OF DIFFERENT METHODS. (FIVE LABELED SAMPLES PER CLASS)

Datasets	Metrics	S-DMM [36]	DCFSL [37]	S3Net [38]	DM-MRN [25]	FSCF-SSL [39]	SPFormer [27]	CTF-SSCL [26]	MambaHSI+ [40]	CGL-PNM [41]	SSEFN [24]	AdaMixCon
Pavia University	Training (s)	419.13	1746.18	2.11	83.71	21966.56	301.08	455.93	371.66	12018.30	21.21	199.22
	Testing (s)	71.38	7.21	6.12	97.98	29.95	9.62	2.14	1.48	82.67	26.04	49.39
	Parameters	33921	4259294	100973	181782	6126910	45225	390357	1630619	5366784	12841160	1320790
Salinas	Training (s)	521.33	1501.26	2.56	92.01	24828.12	102.78	907.11	214.43	25986.77	30.01	335.20
	Testing (s)	93.02	6.88	7.20	119.20	35.53	21.41	3.37	1.02	113.83	41.74	84.92
	Parameters	40385	4270521	164464	183062	6281260	86456	400688	1651618	5384064	16496980	1401926
Houston 2013	Training (s)	181.06	1531.46	3.21	216.04	16261.02	513.12	939.76	1234.25	117972.68	32.03	297.21
	Testing (s)	30.42	5.26	5.56	48.30	21.34	4.52	1.23	3.16	66.56	11.34	27.39
	Parameters	36545	4264360	117851	182102	6188854	46735	394655	1642785	5378880	14341440	1177942

improvement stems from the model's increased complexity, which enables better classification of HSI samples. However, it may cause overfitting on easy samples. The final setup represents our complete framework, achieving the best performance. This demonstrates that DASR stops inference for some easy samples, preventing them from overfitting, and enables the framework to simultaneously classify both easy and hard samples.

TABLE VIII

ABLATION FOR ADAMIXCON ARCHITECTURE ACROSS THREE DATASETS IN TERMS OF OA, AA, AND κ. THE BEST RESULTS ARE SHOWN IN BOLD

Initial	DASR	Refine	Pavia University			Salinas			Houston 2013		
			OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
✓	✗	✗	86.51±4.34	90.75±1.99	82.72±5.11	94.00±1.75	96.32±0.80	93.34±1.93	81.27±1.89	84.04±1.62	79.78±2.04
✗	✗	✓	79.16±3.55	76.66±2.26	73.13±4.06	86.45±8.50	81.54±1.72	84.97±0.96	79.27±2.01	81.73±1.73	77.61±2.18
✓	✗	✓	88.64±3.11	89.58±2.21	85.24±3.91	95.46±0.63	96.89±4.87	94.95±7.01	81.65±2.27	84.26±1.96	80.19±2.45
✓	✓	✓	**90.74±2.40**	**92.78±1.87**	**87.99±2.98**	**95.69±1.00**	**96.90±0.72**	**95.20±1.11**	**83.15±1.86**	**85.56±1.79**	**81.81±2.01**

TABLE IX

ABLATION FOR COMPONENTS IN HSI-MIXCON ACROSS THREE DATASETS IN TERMS OF OA, AA, AND κ. THE BEST RESULTS ARE SHOWN IN BOLD

Inter	SpecScan	SpaScan	Intra	Pavia University			Salinas			Houston 2013		
				OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
✗	✓	✗	✗	89.00±1.83	92.10±1.50	85.78±2.16	94.85±1.21	96.71±0.72	94.27±1.34	82.45±1.88	85.00±1.64	81.06±2.04
✗	✓	✓	✗	89.92±1.50	92.28±1.55	86.91±1.79	95.27±0.68	96.80±0.70	94.74±0.76	82.65±2.34	85.20±2.12	81.27±2.54
✓	✓	✓	✗	90.46±2.25	92.61±1.79	87.62±2.83	95.58±0.99	**97.02±0.66**	95.09±1.10	82.77±2.33	85.34±2.17	81.40±2.53
✓	✓	✓	✓	**90.74±2.40**	**92.78±1.87**	**87.99±2.98**	**95.69±1.00**	96.90±0.72	**95.20±1.11**	**83.15±1.86**	**85.56±1.79**	**81.81±2.01**

TABLE X

ABLATION FOR STRATEGIES IN THE DASR MODULE ACROSS THREE DATASETS IN TERMS OF OA, AA, AND κ. THE BEST RESULTS ARE SHOWN IN BOLD

Local	Global	BvSB	Equal	Pavia University			Salinas			Houston 2013		
				OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
✗	✓	✗	✗	81.02±4.05	78.73±3.17	75.55±4.78	90.13±0.97	88.21±1.95	89.04±1.07	79.72±2.23	82.08±2.08	78.11±2.41
✗	✓	✓	✗	87.28±3.73	90.90±1.80	83.67±4.55	95.54±0.92	96.86±0.77	95.05±1.02	82.89±2.09	85.44±1.86	81.53±2.27
✓	✓	✓	✗	90.38±2.66	92.39±1.70	87.50±3.30	95.64±0.99	**97.02±0.83**	95.15±1.09	82.97±2.10	85.45±1.96	81.62±2.27
✓	✓	✓	✓	**90.74±2.40**	**92.78±1.87**	**87.99±2.98**	**95.69±1.00**	96.90±0.72	**95.20±1.11**	**83.15±1.86**	**85.56±1.79**	**81.81±2.01**

TABLE XI

ABLATION FOR UNCERTAINTY ESTIMATION METRICS IN TERMS OF OA, AA, AND κ. THE BEST RESULTS ARE SHOWN IN BOLD

Metrics	Pavia University			Salinas			Houston 2013		
	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
EP	90.35±2.81	92.64±1.93	87.50±3.46	93.17±3.19	96.74±1.34	92.43±3.52	82.66±1.49	85.20±1.34	81.28±1.62
Max	90.44±2.02	92.41±2.12	87.57±2.58	92.58±3.55	96.58±1.51	91.78±3.52	82.39±2.67	84.77±2.20	81.07±2.88
BvSB	**90.74±2.40**	**92.78±1.87**	**87.99±2.98**	**95.69±1.00**	**96.90±0.72**	**95.20±1.11**	**83.15±1.86**	**85.56±1.79**	**81.81±2.01**

2) Ablation for Components in HSI-MixCon: HSI-MixCon mainly consists of an intersample attention module, a spectral scan, a spatial scan, and an intrasample attention module. Four sets of ablation experiments are designed and conducted across three datasets to evaluate the effectiveness of each component, with results displayed in Table IX.

The first setup employs only a spectral scan but lacks spatial information utilization, resulting in suboptimal performance. The second setup incorporates spectral and spatial scan, achieving improved accuracy, which demonstrates that leveraging both spectral and spatial information facilitates more comprehensive feature extraction from HSI samples. The third setup further introduces the intersample attention module, leading to additional accuracy gains, which verifies that acquiring complementary information from different samples helps to obtain more discriminative features. The final setup integrates an intrasample attention module, attaining the highest accuracy. This proves the effectiveness of the spectral–spatial enhancement after long-range modeling of spectral and spatial scan.

3) Ablation for Strategies in the DASR: The proposed DASR module employs both local and global features to determine the difficulty level of samples. As for the criteria, it jointly utilizes two principles: BvSB and the equal criterion. Four sets of ablation experiments are designed and conducted across three datasets to evaluate the effectiveness of each component, with results displayed in Table X.

The first setup directly refines all the global features output by HSI-MixCon, and the results are poor. This is because the model may overfit the features of easy samples during refinement, leading to feature distortion. The second setup introduces the BvSB criterion to distinguish between easy and hard samples, stopping the inference of easy samples after the first stage, which significantly improves performance and demonstrates the effectiveness of BvSB in discriminating sample difficulty. The third setup further incorporates local features from the first stage, allowing DASR to assess sample difficulty from two complementary perspectives, thereby increasing the reliability of the routing process and further improving accuracy. The final setup adds the equal criterion, which further enhances classification accuracy, proving the validity of the equal criterion.

4) Ablation for Uncertainty Estimation Metrics: BvSB measures prediction uncertainty by calculating the probability gap between the top two predicted classes. A small gap indicates low confidence (ambiguous cases near decision

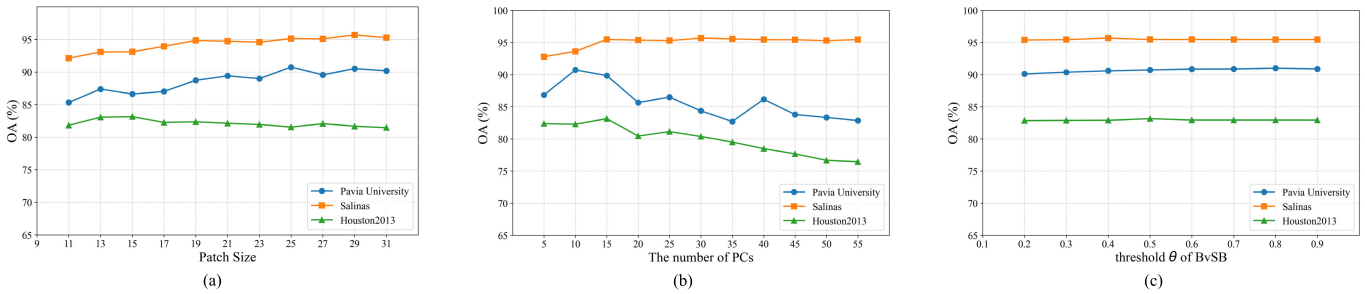

Fig. 7. Classification results of different patch sizes, number of PCs, and threshold θ of BvSB on the Pavia University, Salinas, and Houston 2013 dataset in terms of OA. (a) Patch size, (b) number of PCs, and (c) threshold θ of BvSB.

boundaries), while a large gap reflects high confidence. There are some alternative approaches for uncertainty estimation. The first is predictive entropy (EP) [43], [44], which measures the overall uncertainty in the probability distribution by calculating its information entropy. The second is maximum class probability, which measures confidence simply as the highest predicted probability for any class.

We conduct additional experiments to compare BvSB with these two uncertainty estimation metrics, and the results are shown in Table XI. The results show that BvSB achieves higher accuracy than the other two methods across all three datasets. Entropy measures the overall flatness of the entire probability distribution, while BvSB is less influenced by irrelevant, low-probability classes. Maximum class probability only considers the confidence in the top prediction, but a model can have a moderately high maximum class probability yet still be highly uncertain if the second-best probability is very close. In contrast, BvSB excels at identifying the samples near the decision boundary.

5) Parameter Analysis: To identify the optimal patch size, we evaluate a range of [11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31] and plot the OA values across all three datasets in Fig. 7(a). For Pavia University, the OA reaches its peak at 25 $\times$ 25 pixels. The Salinas dataset demonstrates optimal performance with 29 $\times$ 29 patches. Houston 2013 obtains its highest OA when using 15 $\times$ 15 patches.

To determine the optimal number of principal components (PCs), we test values within the range of [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55] and plot the OA values for all three datasets in Fig. 7(b). For the Pavia University dataset, the optimal number of PCs is 10, beyond which the OA decreases rapidly. Regarding the Salinas dataset, the peak OA occurs at 30 PCs. However, the Houston 2013 dataset performs best with 15 PCs.

To determine the optimal threshold θ in the BvSB criterion, we set the range to [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] and test the OA values across three datasets, as shown in Fig. 7(c). A higher θ imposes stricter requirements for sample routing, which means the difference between the maximum class probability and the second-highest probability of a sample must be sufficiently large for it to be considered high-confidence. For the Pavia University dataset, the optimal θ is 0.8. Regarding the Salinas dataset, the peak OA occurs at 0.4. The Houston 2013 dataset performs best with a threshold of 0.5.

V. Conclusion

In this article, we propose AdaMixCon, an adaptive few-shot hyperspectral image classification framework that achieves enriched spectral–spatial modeling and enables difficulty-aware processing for HSI samples. Our method addresses the challenge of effectively handling samples with rich information and different difficulty levels. It combines sequence modeling with the extraction of intersample and intrasample dependencies to achieve enriched spectral–spatial feature modeling, and adaptively classifies samples based on their difficulty. Extensive evaluations on three public HSI datasets demonstrate that AdaMixCon significantly outperforms state-of-the-art methods, achieving superior classification accuracy.

Acknowledgment

The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] F. Ullah, I. Ullah, R. U. Khan, S. Khan, K. Khan, and G. Pau, "Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3878–3916, 2024.

[2] O. Akar and E. Tunc Gormus, "Land use/land cover mapping from airborne hyperspectral images with machine learning algorithms and contextual information," *Geocarto Int.*, vol. 37, no. 14, pp. 3963–3990, Jul. 2022.

[3] V. P. Yele, R. R. Sedamkar, and S. Alegavi, "Systematic analysis of effective segmentation and classification for land use land cover in hyperspectral image using deep learning methods: A review of the state of the art: Reviewing deep learning techniques for land use and cover in hyperspectral images," in *Proc. 20th CSI Int. Symp. Artif. Intell. Signal Process. (AISP)*, Feb. 2024, pp. 1–8.

[4] R. Rajabi, A. Zehtabian, K. D. Singh, A. Tabatabaeenejad, P. Ghamisi, and S. Homayouni, "Editorial: Hyperspectral imaging in environmental monitoring and analysis," *Frontiers Environ. Sci.*, vol. 11, Jan. 2024, Art. no. 1353447.

[5] M. B. Stuart, M. Davies, M. J. Hobbs, T. D. Pering, A. J. S. McGonigle, and J. R. Willmott, "High-resolution hyperspectral imaging using low-cost components: Application within environmental monitoring scenarios," *Sensors*, vol. 22, no. 12, p. 4652, Jun. 2022.

[6] A. Nisha and A. Anitha, "Current advances in hyperspectral remote sensing in urban planning," in *Proc. 3rd Int. Conf. Intell. Comput. Instrum. Control Technol. (ICICICT)*, Aug. 2022, pp. 94–98.

[7] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[8] A. Özdemir and K. Polat, "Deep learning applications for hyperspectral imaging: A systematic review," *J. Inst. Electron. Comput.*, vol. 2, no. 1, pp. 39–56, 2020.

[9] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.
[10] W. Wei, L. Tong, B. Guo, J. Zhou, and C. Xiao, "Few-shot hyperspectral image classification using relational generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
[11] J. Zhao, J. Zhang, H. Huang, and J. Zhang, "Enhancing semi-supervised few-shot hyperspectral image classification via progressive sample selection," *Remote Sens.*, vol. 16, no. 10, p. 1747, May 2024.
[12] M. Zhang, H. Liu, M. Gong, H. Li, Y. Wu, and X. Jiang, "Cross-domain self-taught network for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023.
[13] Y. Dong, B. Zhu, X. Yang, and X. Ma, "Deep metric learning based on Brownian covariance representation for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.
[14] H. Tang, C. Zhang, D. Tang, X. Lin, X. Yang, and W. Xie, "Few-shot hyperspectral image classification with deep fuzzy metric learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025.
[15] H. Wu, Z. Xue, S. Zhou, and H. Su, "Overcoming granularity mismatch in knowledge distillation for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–17, 2025.
[16] W. Li et al., "Few-shot learning based on embedded self-distillation and adaptive Wasserstein distance for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.
[17] Z. Ye, J. Wang, T. Sun, J. Zhang, and W. Li, "Cross-domain few-shot learning based on graph convolution contrast for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.
[18] H. Liu, J. He, Y. Li, and Y. Bi, "Multilevel prototype alignment for cross-domain few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.
[19] Q. Zhu, H. Li, W. Deng, Q. Guan, and J. Luo, "From intra-distinctiveness to inter-invariance: A cycle-resemblance few-shot transformation network for cross-domain hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, 2025.
[20] Z. Zhang, D. Gao, D. Liu, and G. Shi, "Spectral–spatial domain attention network for hyperspectral image few-shot classification," *Remote Sens.*, vol. 16, no. 3, p. 592, Feb. 2024.
[21] R. Fan, L. Tong, J. Zhou, B. Guo, and C. Xiao, "Prototypical network with residual capsule for few-shot hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[22] J. Sheng, J. Zhou, J. Wang, P. Ye, and J. Fan, "DualMamba: A lightweight spectral–spatial Mamba-convolution network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.
[23] X. Xu and Z. Lin, "MixCon: A hybrid architecture for efficient and adaptive sequence modeling," in *Proc. Eur. Conf. Artif. Intell.*, 2024, pp. 1027–1034.
[24] S. Liu, C. Fu, Y. Duan, X. Wang, and F. Luo, "Spatial–spectral enhancement and fusion network for hyperspectral image classification with few labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.
[25] J. Zeng, Z. Xue, L. Zhang, Q. Lan, and M. Zhang, "Multistage relation network with dual-metric for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.
[26] B. Xi, Y. Zhang, J. Li, Y. Li, Z. Li, and J. Chanussot, "CTF-SSCL: CNN-transformer for few-shot hyperspectral image classification assisted by semisupervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5532617.
[27] Z. Li, Z. Xue, Q. Xu, L. Zhang, T. Zhu, and M. Zhang, "SPFormer: Self-pooling transformer for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–19, 2024.
[28] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[29] Y. Qing, W. Liu, L. Feng, and W. Gao, "Improved transformer net for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 11, p. 2216, Jun. 2021.
[30] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "Mambahsi: Spatial–spectral mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
[31] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[32] F. Xiao, H. Xiang, C. Cao, and X. Gao, "Neural architecture search-based few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.
[33] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
[34] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 103031–103063.
[35] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
[36] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.
[37] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
[38] Z. Xue, Y. Zhou, and P. Du, "S3Net: Spectral–spatial Siamese network for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
[39] Z. Li et al., "Few-shot hyperspectral image classification with self-supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.
[40] Y. Wang, L. Liu, J. Xiao, D. Yu, Y. Tao, and W. Zhang, "MambaHSI$: Multidirectional state propagation for efficient hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.
[41] W. N. Khotimah, M. Bennamoun, F. Boussaid, L. Xu, and F. Sohel, "Dual-phase framework for few-shot hyperspectral image classification with spatiospectral masked autoencoder and episode training," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, 2025.
[42] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7260–7268.
[43] H. Wang and L. Wang, "Collaborative active learning based on improved capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–26, 2023.
[44] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

Chenchen Wang received the B.E. degree from the College of Future Information Technology, Fudan University, Shanghai, China, in 2024, where he is currently pursuing the M.E. degree.

His main research interests include computer vision and hyperspectral analysis.

Jiamu Sheng received the B.E. degree from the College of Future Information Technology, Fudan University, Shanghai, China, in 2022, and the M.E. degree from the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, in 2025.

His main research interests include computer vision, image quality assessment, and hyperspectral analysis.

Jingyi Zhou (Student Member, IEEE) received the B.E. degree from the College of Future Information Technology, Fudan University, Shanghai, China, in 2023, where she is currently pursuing the M.E. degree.

Her main research interests include computer vision, hyperspectral analysis, sentiment analysis, and depth estimation.

Zhende Song received the B.E. degree from Shanghai University, Shanghai, China, in 2022. He is currently pursuing the M.S. degree with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai.

He has published papers in reputable journals and conferences, such as ACM'MM. His research interests include computer vision, computer graphics, and machine learning.

Jiayuan Fan (Senior Member, IEEE) received the Ph.D. degree in information engineering from Nanyang Technological University, Singapore, in 2015.

After her graduation, she was a Research Scientist at the Institute for Infocomm Research, A*STAR, Singapore. She is currently an Associate Professor with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China. Her main research interests include computer vision, image forensic analysis, and applications.

Zhouchen Lin (Fellow, IEEE) received the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 2000.

He is currently a Boya Special Professor with the State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University. He has published over 360 technical articles, collecting more than 42 000 Google Scholar citations. His research interests include machine learning and numerical optimization.

Dr. Lin is a member of the ICML Board of Directors. He is a fellow of IAPR, AAIA, CCF, and CSIG. He has been the Program Co-Chair of ICPR 2022 and the Area Chair or the Senior Area Chair of ACML, ACCV, CVPR, ICCV, NIPS/NeurIPS, AAAI, IJCAI, ICLR, and ICML for many times. He is the Associate Editor-in-Chief of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.